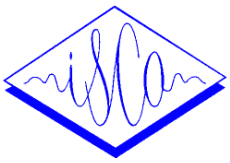

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

Real-Time Telephone Transmission Simulation for Speech Recognizer and Dialogue System Evaluation and Improvement

Sebastian Möller[*], *Hervé Bourlard*[**]

[*]Institute of Communication Acoustics (IKA), Ruhr-University Bochum, D-44780 Bochum, Germany
Tel. +49 234 322 3979, Fax +49 234 321 4165, E-mail: moeller@ika.ruhr-uni-bochum.de

[**]IDIAP – Institut Dalle Molle d'Intelligence Artificielle Perceptive, CP 592, CH-1920 Martigny, Switzerland
Tel. +41 27 721 7711, Fax +41 27 721 7712, E-mail: bourlard@idiap.ch

ABSTRACT

Recognizer performance in telephone-based spoken dialogue systems may be strongly affected by the transmission channel. In order to investigate the impact of different parts of the transmission channel in more detail, a simulation model is presented. It implements all transmission characteristics of modern telephone networks, based on instrumentally measurable values as they are used by network planners. The simulation shows real-time capability and runs on a programmable DSP-based hardware. It can be used for a systematic investigation of recognizer performance as a function of transmission channel degradations, for producing training material with specified transmission characteristics, or for estimating the impact of transmission impairments on dialogue flow and system usability. The impact of transmission channel characteristics on the performance of a speech recognizer integrated in an interactive voice server is analyzed in more detail. It turns out that specific transmission characteristics may lead to a recognition degradation which otherwise would not have been expected from the standard training material. An outlook is given on future extensions of the simulation model, in order to better cover effects of mobile and IP-based telephone systems.

1. INTRODUCTION

The quality of spoken dialogue systems operated over the wired or mobile telephone network is affected by the transmission channel. On one hand, transmission impairments may lead to a considerable degradation of recognition performance, and consequently to problems in speech understanding and dialogue flow. On the other hand, the channel transfer characteristics will impact the sound quality of the speech output, be it synthesized or naturally produced. Both effects have to be taken into consideration when the quality of spoken dialogue systems is the focus of interest, because they influence conversation quality, system usability, and finally the acceptability of voice-operated services.

Transmission impairments which are found in telephony may be very diverse in nature. Due to the changes in the growing and liberalized telecommunication market, impairments which have been common in traditional – analogue as well as digital – networks (e.g. loss, circuit noise, quantizing distortion) are accompanied by echoes, delay, and non-linear distortions resulting from low-bitrate codecs. Mobile handsets show transfer characteristics which are very different from those of traditionally shaped handsets, and may favorize the pick-up of ambient noise. Additional time-variant impairments arise in mobile and packet-based networks, and from signal processing devices (e.g. echo cancellers). As a combination of all such impairments will occur in a real telephone connection, the joint effects have to be taken into account when the impact of the transmission channel on system performance is to be estimated (cf. Wyard, 1993).

While different types of models have been established predicting the joint effects of transmission impairments on speech communication quality as perceived by a human listener (comparative instrumental models, network planning models; cf. e.g. Möller, 2000, for a review), no corresponding approach exists for the quality degradation to be expected in speech recognition. In order to assess or reduce the impact of the transmission channel on speech recognition, two approaches have often been followed up to now: either the effects of single transmission impairments (e.g. circuit noise, codecs) on the recognition have been studied separately (e.g. Euler and Zinke, 1993), or expensive large telephone speech databases have been collected in order to appropriately cover real-life impairments in the recognizer's training material.

In this paper we present a different approach. A real-time telephone line simulation system was developed at IKA, Ruhr-University Bochum (Möller, 2000). It permits all relevant input parameters to be set in a controlled way. These parameters are planning values of traditional as well as of new (mobile, IP-based) networks. The set-up and the characteristics of the simulation system are discussed in more detail in the next section. It is used to analytically investigate the effects of different types of transmission degradations on the speech recognition component of a spoken dialogue system. The experimental conditions and the obtained recognition results are discussed in Sections 3 and 4. A final discussion (Section 5) identifies desirable extensions of the simulation model.

2. TELEPHONE TRANSMISSION SIMULATION

In the planning process of spoken dialog system based telephone services, it is in general not predictable under which transmission conditions the service will be used. Although it

may be obvious that services operated by mobile network companies will mainly be accessed through a mobile channel, or that e.g. traffic information systems will often be used by car drivers in a noisy environment, it is highly desirable to have systems which are operational under an even wider range of transmission channel and environmental characteristics. As the actual connection characteristics are not in the hands of the system developer, planning values are appropriate parameters in the developmental phase of new services. This assumption is in contrast to work carried out by Tarcisio et al. (1999). Although it is possible for the model presented here to simulate connections with specific components, this is not a necessary prerequisite of our approach.

The basis for our simulation system is a planning configuration defined by the International Telecommunication Union (ITU-T) in order to estimate end-to-end transmission performance (ITU-T Rec. G.107, 2000). This configuration considers instrumentally measurable transmission parameters of terminal, switching and connection elements used in modern – analogue as well as digital – telephone networks. Specific time-variant characteristics of mobile and packet-based networks (transmission errors, fading, etc.) are not yet taken into account. Transfer characteristics of the individual network paths are characterized by so-called loudness ratings, i.e. scalar planning values which reflect the sensitivity of the human ear to a certain extent. The same principle is applied to different noise sources, which are represented by psophometrically or A-weighted power levels. Whereas waveform coders are characterized by the amount of quantizing noise they introduce, the perceptively relevant effects of non-waveform low-bitrate coders are taken into account using an integral quality degradation factor inherent to the specific coder-decoder pair, the so-called equipment impairment factor. Details on all planning values can be found in the ITU-T literature (e.g. ITU-T Rec. G.107, 2000).

Starting from this basic planning configuration of a telephone connection, it is desirable to implement a flexible simulation model, where all transmission parameters can be adjusted in a controlled way, and within the ranges which can be expected in real networks. Real-time capability allows the model to be used in real conversations, e.g. for assessing the impact of transmission impairments on dialogue flow. For processing recognition data, it is more convenient to provide an off-line version. On the side of the human user, the simulation should be terminated with a piece of terminal equipment typical for the application, e.g. a standard or mobile handset or a hands-free terminal.

Following these requirements, a programmable DSP hardware system was chosen for model implementation. It works in real-time, thus enabling realistic conversations as well as off-line data processing. The following telephone line degradations can be modeled:

- attenuation and frequency distortion of the transmission channel (send loudness ratings, *SLR*, and receive loudness ratings, *RLR*)
- continuous (white) circuit noise representing all potentially distributed noise sources, both on the channel (*Nc*, narrow-band) and at the receive side (*Nfor*, wideband)
- transmission channel bandwidth (BP with 300-3400 Hz or 50-7000 Hz)
- different speech codecs: several ITU-T codecs, as well as a North American cellular (IS-54) and proprietary codecs
- quantizing distortion due to waveform codecs (log. PCM) or D/A-A/D conversions
- ambient room noise at the send (*Ps*) and receive side (*Pr*)
- absolute delay (*Ta*)
- talker echo with one-way delay *T* and attenuation *Le*
- listener echo with round-trip delay *Tr* and attenuation *WEPL*
- listener sidetone with attenuation *Lst*

The implemented structure is depicted in Figure 1. The triangles represent FIR filters or programmable attenuators, the rectangles delay lines (*T*, *Ta* and *Tr*), codecs, and the channel bandpass filter (BP). Several ITU-T standardized codecs are implemented on specific separate DSP boards. Three such boards are available to us, so it is possible to simulate asynchronous codec tandems with up to three coding-decoding processes. It can be seen from Figure 1 that all relevant speech paths (the main transmission path, the talker sidetone path, talker and listener echo paths) are modeled independently, in a full-duplex mode (indices 1 and 2). A difference to real-life networks is the implementation of the talker echo path: whereas the talker echo passes through two coding-decoding processes (the reflections normally take place at the far end hybrid), the simulation only takes into account one coding-decoding process. This compromise was made for stability reasons, due to the feedback loop which otherwise would have been formed between the two talker echo paths.

The simulation model can be connected to different types of terminal equipment, in a four-wire mode. For the experiments reported below, standard handsets were used. These handsets are characterized by their sensitivity in sending and receiving direction (expressed by the filter characteristics *SLRset* and *RLRset*), and their difference in sensitivity between direct (speech) sound and diffuse (e.g. ambient room noise) sound. Additional filters (*SLR'*, *RLR'*) permit the whole sending (*SLR = SLRset * SLR'*) and receiving (*RLR = RLRset * RLR'*) characteristic to be adjusted to a desired shape. Ambient room noise can be inserted at the speaker's or listener's side, via additional loudspeakers in the test cabinets. When the effects of room noise are to be taken into account, it is important to put the human communication partner into a noisy environment, in order to enclose speaking style variations (Lombard reflex) in the test set-up.

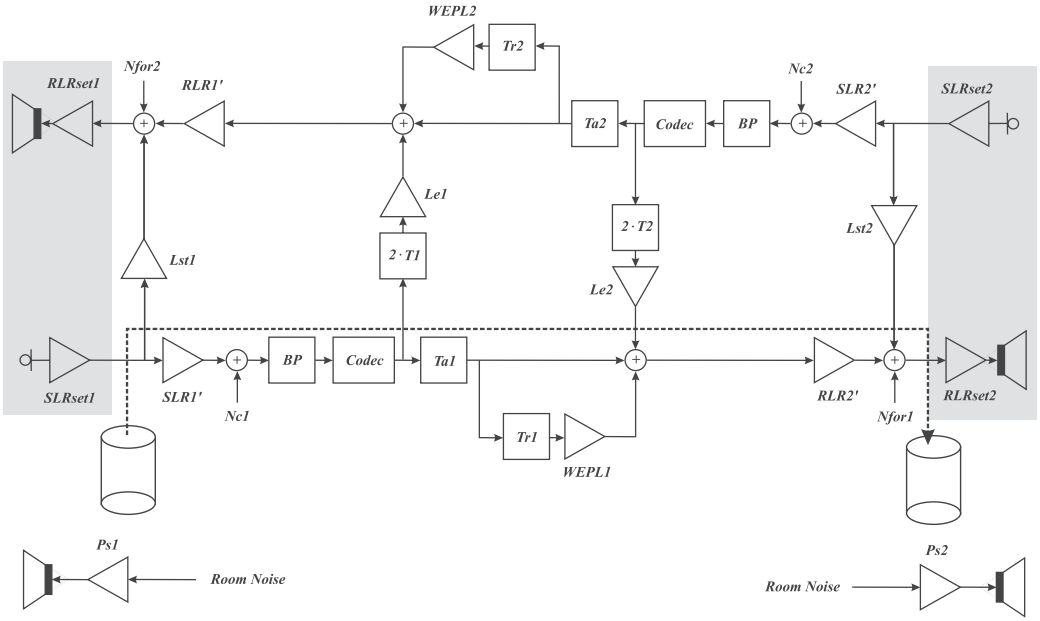

Figure 1: Configuration of the telephone line simulation model. Gray boxes represent terminal equipment, e.g. handsets, the dashed line shows the transmission set-up for Section 3. All parameters are discussed in the text.

The system presented above has originally been implemented in order to assess the effects of transmission impairments on speech communication quality between humans. For that aim, both sides can be connected to standard terminal equipment. For investigating the impact of the transmission channel on a spoken dialogue system, the latter can be connected to one side of the transmission line, replacing one of the gray boxes. The *SLR'* and *RLR'* filters at that side may then be used to model the transmission characteristics between the system connection and the network reference point. In this case, room noise at the dialogue system's side will not be inserted, and *Lst* will be set to infinity.

3. APPLICATION TO SPEECH RECOGNIZER ASSESSMENT

The system's capabilities make it a very useful tool for three different applications: (1) to assess the impact of joint transmission impairments on the performance of a speech recognizer, (2) to produce a large amount of recognizer training material with defined transmission characteristics, thus multiplying the available data amount, and (3) its real-time capability allows the effects of transmission impairments on the course of the dialogue to be assessed, including speech understanding and dialogue managing strategies. In the primary study reported here, we investigated the first point in more detail.

Our test object is a continuous speech recognizer which is part of an interactive voice server. This server provides information on the restaurants of the city of Martigny, Switzerland, integrating voice and internet access. It has been implemented as a part of the Swiss CTI-funded InfoVOX project. At the time the experiments have been carried out, only a first non-optimized prototype version of the recognizer was available. Thus, the overall level of recognition results will reflect the developmental phase they have been obtained in.

The large-vocabulary continuous Swiss-French speech recognizer is a hybrid HMM/ANN system. ANN weights, HMM phone models and phone prior probabilities have been trained on the Swiss-French PolyPhone Database (Chollet et al., 1996), using 4,293 information service calls (2,407 female, 1,886 male speakers) which have been collected over the Swiss telephone network. The recognizer's dictionary was built from 255 initial Wizard-of-Oz dialogue transcriptions on the restaurant information task. The same transcriptions were used to set up 2-gram and 3-gram language models. Log-RASTA feature coefficients consisted of 12 MFCC coefficients, 12 derivations, plus energy and energy derivations, using a 10th order LPC analysis and 17 critical band filters.

Table 1: Telephone line transmission parameters. All other parameters have been set to their default values, as given in ITU-T Rec. G.107 (2000).

#	*SLR'* (dB)	*RLR'* (dB)	*Nc* (dBm0p)	codec	note
A	8	2	-70	G.711	default
B	8	2	-55	G.711	low noise
C	8	2	-40	G.711	high noise
D	13	7	-70	G.711	quiet conn.
E	8	2	-70	G.726 (32kbit/s)	DCME
F	8	2	-70	G.728 (16kbit/s)	DECT
G	8	2	-70	G.729 (8kbit/s)	
H	8	2	-70	IS-54	NA cellular
I	8	2	-55	IS-54	noise+cell.
J	8	2	-40	IS-54	noise+cell.
K	8	2	-70	MNRU, Q=20dB	sig. corr. n.
L	8	2	-70	MNRU, Q=10dB	sig. corr. n.

As test data, Wizard-of-Oz-based dialogue utterances were collected from 10 different speakers (6m, 4f) and then processed by the model described above, simulating telephone lines with defined characteristics. 15 test sentences were solicited from each speaker, which were similar (though not identical) to the WoZ transcriptions, and which contained each at least two 'field' values for the dialogue flow. Speakers read the utterances aloud in a natural way. The utterances were recorded, transmitted through the simulation model, and then used as input to the recognizer. The exact transmission parameter settings are given in Table 1. They include a default ISDN connection (A), different levels of circuit noise (B, C), a 'quiet' connection (D), different types of codecs (E-H), combinations of noise with a cellular codec (I, J), as well as two levels of signal correlated noise (K, L), inserted with a Modulated Noise Reference Unit (MNRU) at the position of the codec. These transmission impairments cover a wide range of perceptively very different degradations.

4. RECOGNITION RESULTS AND DISCUSSION

Recognition results have been scored in two ways. In the first part, mean values for the percentage of correct words (corr), of substitutions (sub), deletions (del) and insertions (ins), as well as of all errors (err) have been calculated for each connection setting. In the second part, the same values have been calculated, but only for the 224 keywords which are actually used by the speech understanding component of the system (total vocabulary size: 939 words).

Table 2: Mean recognition results.

#	mean results, all words					mean results, keywords only				
	corr	sub	del	ins	err	corr	sub	del	ins	err
-	57.4	28.6	14.0	3.7	46.4	69.5	10.7	19.8	4.2	34.7
A	47.2	31.1	21.7	3.2	56.0	59.5	12.2	28.3	4.8	45.3
B	39.7	30.7	29.7	2.3	62.7	53.5	13.8	32.7	3.5	50.0
C	10.4	19.3	70.3	0.3	89.8	14.0	12.0	74.0	0.2	86.2
D	34.8	32.9	32.3	1.8	66.9	47.7	16.7	35.7	2.7	55.0
E	41.7	31.7	26.6	2.3	60.5	53.3	12.5	34.2	4.7	51.3
F	45.0	30.4	24.6	2.6	57.6	57.0	13.5	29.5	4.3	47.3
G	46.7	35.9	17.4	3.8	57.0	57.8	16.7	25.5	7.0	49.2
H	45.9	36.1	18.0	4.5	58.6	58.0	17.8	24.2	6.3	48.3
I	42.0	37.5	20.4	3.0	61.0	53.2	19.2	27.7	7.2	54.0
J	17.5	35.3	47.1	0.9	83.3	27.8	21.8	50.3	2.7	74.8
K	48.7	28.3	23.0	1.9	53.2	61.2	10.8	28.0	3.8	42.7
L	17.5	22.2	60.3	0.2	82.7	22.5	8.5	69.0	0.8	78.3

Both sets of results are listed in Table 2, for the original data (-) and for each of the telephone line settings A-L. It turns out that already a standard ISDN connection (A) considerably decreases the recognition performance (~10% for both all words and keywords). This may be due to a lower SNR for the transmitted data (also reflected in condition D), as well as to the strict bandwidth limitation (300-3400 Hz) to be expected in long-distance or international calls. The overall recognition level is relatively low, which can be explained by the non-optimized prototype version which was available during that phase of the project, as well as by a general mismatch between training and test conditions (no training on self-collected data). It can be expected that a better training of the language model as well as a better dictionary will increase the performance. Further improvements can be expected when using a context-free grammar.

The recognizer using RASTA coefficients shows a strong noise sensitivity, both with and without additional codecs (B, C, I, J). On the other hand, it is nearly completely insensitive to low-bitrate codecs (also cellular) and to moderate levels of signal-correlated noise (E-H, K). An exception is the G.726 ADPCM codec, for which recognition performance drops about 6%. Strong signal-correlated noise strongly impacts recognition.

5. CONCLUSION

The exemplary recognizer evaluation shows that the telephone line simulation model presented here can fruitfully be used for evaluating and optimizing spoken dialogue system components with respect to the influences of the transmission channel. Dependencies may be uncovered which otherwise would not have been obvious, and which might have lead to strong impacts in later system development phases.

For the future, it would be useful to extend the simulation model in order to better capture the (time-variant) impairments typical for mobile as well as IP-based networks. Further simulation will be necessary when services are accessed via hands-free terminals and with strong ambient noise backgrounds, e.g. in car environments.

ACKNOWLEDGEMENTS

This study was supported by the EU-funded project SPeech and HEARing (SPHEAR). The dialogue system was developed in the Swiss-funded CTI project InfoVOX. The authors would like to thank the members of IDIAP for their support.

REFERENCES

Chollet, G., Cochard, J.-L., Constantinescu, A., Jaboulet, C., Langlais, Ph., "*Swiss French PolyPhone and PolyVar: Telephone Speech Databases to Model Inter- and Intra-Speaker Variability*", Technical Report RR-96-01, IDIAP, CH-Martigny, 1996.

Euler, S., Zinke, J., "*The Influence of Speech Coding Algorithms on Automatic Speech Recognition*", Proc. ICASSP'94, AUS-Adelaide, 1994, pp. I-621-624.

ITU-T Recommendation G.107, "*The E-Model, a Computational Model for Use in Transmission Planning*", International Telecommunication Union, CH-Geneva, 2000.

Möller, S., "*Assessment and Prediction of Speech Quality in Telecommunications*", Kluwer Academic Publishers, USA-Boston, 2000.

Tarcisio, C., Daniele, F., Roberto, G., Marco, O., "*Use of Simulated Data for Robust Telephone Speech Recognition*", Proc. EUROSPEECH'99, H-Budapest, 1999, pp. 2825-2828.

Wyard, P., "*The Relative Importance of the Factors Affecting Recognizer Performance with Telephone Speech*", Proc. EUROSPEECH'93, D-Berlin, 1993, pp. 1805-1808.